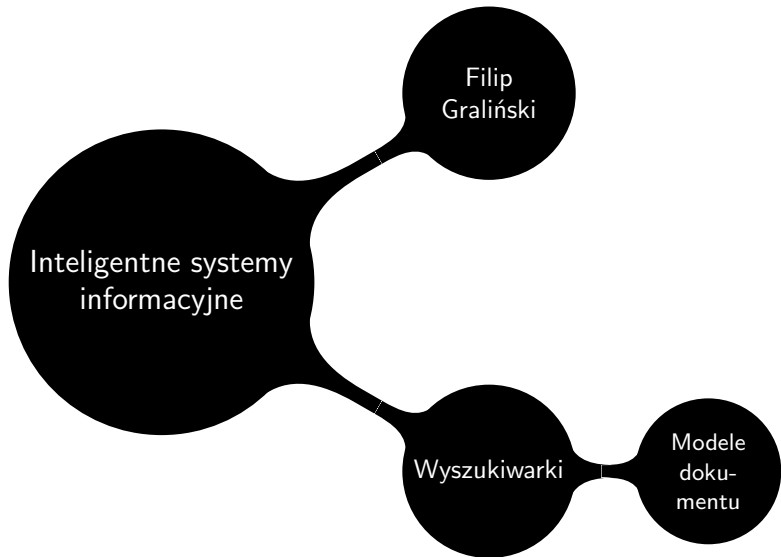


```
graph LR; A[Inteligentne systemy informacyjne] --- B[Filip Graliński]; A --- C[Wyszukiwarki]
```

Inteligentne systemy  
informacyjne

Filip  
Graliński

Wyszukiwarki



## odwrócony indeks

doc1		<i>Ala ma kota.</i>
doc2		<i>Podobno jest kot w butach.</i>
doc3		<i>Ty chyba masz kota!</i>
doc4		<i>But chyba zgubiłem.</i>

## odwrócony indeks

doc1		<i>Ala ma kota.</i>
doc2		<i>Podobno jest kot w butach.</i>
doc3		<i>Ty chyba masz kota!</i>
doc4		<i>But chyba zgubiłem.</i>

*stop words: jest, w*

## odwrócony indeks

doc1		<i>Ala ma kota.</i>
doc2		<i>Podobno jest kot w butach.</i>
doc3		<i>Ty chyba masz kota!</i>
doc4		<i>But chyba zgubiłem.</i>

*stop words: jest, w*

<i>ala</i>	→	1
<i>but</i>	→	2 4
<i>chyba</i>	→	3 4
<i>kot</i>	→	1 2 3
<i>mieć</i>	→	?
<i>podobno</i>	→	2
<i>ty</i>	→	3
<i>zgubić</i>	→	4

## zapytania boole'owskie

- ▶ *pizzeria Poznań dowóz*

## zapytania boole'owskie

- ▶ *pizzeria AND Poznań AND dowóz*

## zapytania boole'owskie

- ▶ *pizzeria AND Poznań AND dowóz*
- ▶ *(pizzeria OR pizza OR tratoria) AND Poznań AND dowóz*



## zapytania boole'owskie

- ▶ *pizzeria AND Poznań AND dowóz*
- ▶ *(pizzeria OR pizza OR tratoria) AND Poznań AND dowóz*
- ▶ *pizzeria AND Poznań AND dowóz AND NOT golonka*

## zapytania boole'owskie

- ▶ *pizzeria AND Poznań AND dowóz*
- ▶ *(pizzeria OR pizza OR tratoria) AND Poznań AND dowóz*
- ▶ *pizzeria AND Poznań AND dowóz AND NOT golonka*

### Jak domyślnie interpretować zapytanie?

- ▶ jako zapytanie AND – być może za mało dokumentów
- ▶ rozwiązanie pośrednie?
- ▶ jako zapytanie OR – być może za dużo dokumentów

## dokument jako wektor

doc1		<i>Ala ma kota.</i>
doc2		<i>Podobno jest kot w butach.</i>
doc3		<i>Ty chyba masz kota!</i>
doc4		<i>But chyba zgubiłem.</i>

## dokument jako wektor

doc1	<i>Ala ma kota.</i>
doc2	<i>Podobno jest kot w butach.</i>
doc3	<i>Ty chyba masz kota!</i>
doc4	<i>But chyba zgubiłem.</i>

	<i>ala</i>	<i>but</i>	<i>chyba</i>	<i>kot</i>	<i>mieć</i>	<i>podobno</i>	<i>ty</i>	<i>zgubić</i>
doc1	1	0	0	1	1	0	0	0
doc2	0	1	0	1	0	1	0	0
doc3	0	0	1	1	1	0	1	0
doc4	0	1	1	0	0	0	0	1

## dokument jako wektor

doc1	<i>Ala ma kota.</i>
doc2	<i>Podobno jest kot w butach.</i>
doc3	<i>Ty chyba masz kota!</i>
doc4	<i>But chyba zgubiłem.</i>
doc5	<i>Kot ma kota.</i>

	<i>ala</i>	<i>but</i>	<i>chyba</i>	<i>kot</i>	<i>mieć</i>	<i>podobno</i>	<i>ty</i>	<i>zgubić</i>
doc1	1	0	0	1	1	0	0	0
doc2	0	1	0	1	0	1	0	0
doc3	0	0	1	1	1	0	1	0
doc4	0	1	1	0	0	0	0	1

## dokument jako wektor

doc1	<i>Ala ma kota.</i>
doc2	<i>Podobno jest kot w butach.</i>
doc3	<i>Ty chyba masz kota!</i>
doc4	<i>But chyba zgubiłem.</i>
doc5	<i>Kot ma kota.</i>

	<i>ala</i>	<i>but</i>	<i>chyba</i>	<i>kot</i>	<i>mieć</i>	<i>podobno</i>	<i>ty</i>	<i>zgubić</i>
doc1	1	0	0	1	1	0	0	0
doc2	0	1	0	1	0	1	0	0
doc3	0	0	1	1	1	0	1	0
doc4	0	1	1	0	0	0	0	1
doc5	0	0	0	?	1	0	0	0

jak uwzględnić frekwencję wyrazu?

$tf_{t,d}$

jak uwzględnić frekwencję wyrazu?

$$tf_{t,d}$$

$$1 + \log(tf_{t,d})$$



jak uwzględnić frekwencję wyrazu?

$$tf_{t,d}$$

$$1 + \log(tf_{t,d})$$

$$\begin{cases} 1, & \text{jeśli } tf_{t,d} > 0 \\ 0, & \text{w przeciwnym razie} \end{cases}$$

jak uwzględnić frekwencję wyrazu?

$$tf_{t,d}$$

$$1 + \log(tf_{t,d})$$

$$\begin{cases} 1, & \text{jeśli } tf_{t,d} > 0 \\ 0, & \text{w przeciwnym razie} \end{cases}$$

$$0.5 + \frac{0.5 \times tf_{t,d}}{\max_t(tf_{t,d})}$$

## odwrotna częstość w dokumentach

Czy wszystkie wyrazy są tak samo ważne?

## odwrotna częstość w dokumentach

Czy wszystkie wyrazy są tak samo ważne?

**NIE.** Wyrazy pojawiające się w wielu dokumentach są mniej ważne.

## odwrotna częstość w dokumentach

Czy wszystkie wyrazy są tak samo ważne?

**NIE.** Wyrazy pojawiające się w wielu dokumentach są mniej ważne.

Aby to uwzględnić, przemnażamy frekwencję wyrazu przez **odwrotną częstość w dokumentach** (*inverse document frequency*):

$$\text{idf}_t = \log \frac{N}{\text{df}_t}$$

,

$\text{idf}_t$  – odwrotna częstość wyrazu  $t$  w dokumentach

$N$  – liczba dokumentów w kolekcji

$\text{df}_f$  – w ilu dokumentach wystąpił wyraz  $t$ ?

## odwrotna częstość w dokumentach (cd.)

wyraz wystąpił...	$idf_t = \dots$
w 1 dokumencie	

## odwrotna częstość w dokumentach (cd.)

wyraz wystąpił...	$\text{idf}_t = \dots$
w 1 dokumencie	$= \log N/1 = \log N$

## odwrotna częstość w dokumentach (cd.)

wyraz wystąpił...	$\text{idf}_t = \dots$
w 1 dokumencie	$= \log N/1 = \log N$
2 razy w kolekcji	$= \log N/2$ lub $\log N$



## odwrotna częstość w dokumentach (cd.)

wyraz wystąpił...	$idf_t = \dots$
w 1 dokumencie	$= \log N/1 = \log N$
2 razy w kolekcji	$= \log N/2$ lub $\log N$
w połowie dokumentów	$= \log N/(N/2) = \log 2$

## odwrotna częstość w dokumentach (cd.)

wyraz wystąpił...	$idf_t = \dots$
w 1 dokumencie	$= \log N/1 = \log N$
2 razy w kolekcji	$= \log N/2$ lub $\log N$
w połowie dokumentów	$= \log N/(N/2) = \log 2$
we wszystkich dokumentach	$= \log N/N = \log 1 = 0$

## odwrotna częstość w dokumentach (cd.)

wyraz wystąpił...	$idf_t = \dots$
w 1 dokumencie	$= \log N/1 = \log N$
2 razy w kolekcji	$= \log N/2$ lub $\log N$
w połowie dokumentów	$= \log N/(N/2) = \log 2$
we wszystkich dokumentach	$= \log N/N = \log 1 = 0$

Zamiast  $tf_{t,d}$  będziemy w wektorach rozpatrywać wartości:

$$tf-idf_{t,d} = tf_{t,d} \times idf_t$$

## odwrotna częstość w dokumentach (cd.)

wyraz wystąpił...	$idf_t = \dots$
w 1 dokumencie	$= \log N/1 = \log N$
2 razy w kolekcji	$= \log N/2$ lub $\log N$
w połowie dokumentów	$= \log N/(N/2) = \log 2$
we wszystkich dokumentach	$= \log N/N = \log 1 = 0$

Zamiast  $tf_{t,d}$  będziemy w wektorach rozpatrywać wartości:

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t$$

*Overlap score measure:*

$$\text{Score}(q, d) = \sum_{t \in q} \text{tf-idf}_{t,d}$$

## Opowieść A

*Dziewczyna w akademiku kupiła weza, z czasem podrosł. Zrzucone skory na ścianach, rysunki, wycinki, waz slizgał sie luzem, sypał w lozku, itd. w pewnym momencie waz przestal jesc myszy, rozlozyl sie i zaczął "nadymac". Telefony do sklepu, weterynarza i kogostam jeszcze nie daly skutku i nikt nie potrafil doradzić dziewczynie jak to ogarnac. Waz lezal i sie "nadymal", a dziewczyna zamartwiala sie przekonana, ze choruje. Zadzwonila przypadkiem do zoo, opowiedziala cala sytuacje i uslyszala: Proszę wyjsc z pokoju i zamknac drzwi - waz jest zdrowy, a na lozku lezy i sie "nadyma" bo sprawdza czy jest w stanie pania zmiescic...*

## Opowieść B

*Anegdota o wężu Historia wydarzyła się koleżance. Dziewczyna hodowała pytona (pytony nie mają kłów jadowych, czy jak tam to się nazywa) . Wąż był już dość spory i nie trzymała go w żadnym specjalnym terrarium tylko luzem w mieszkaniu. Wąż snuł się w po domu, wylegiwał na meblach i parapetach. Ale pewnego dnia przestał jeść. Po kilku dniach jego głodówki zaniepokojona właścicielka zabrała go do weterynarza. Ten obejrzał i powiedział, że wygląda na zdrowego. Koleżanka zabrała go do chaty, ale sytuacja jeszcze przez kilka dni się utrzymywała - dalej nie jadł. Pewnej nocy obudziła się i zobaczyła go obok siebie na łóżku swojego węża wyciągniętego wzdłuż niej jak kij od miotły. W pierwszej chwili przestraszyła się, że zdechł, ale okazało się, że żyje. Następnego dnia zabrała go znowu do weterynarza i opowiedziała, że dalej nie je, a poza tym wspomniała o dziwnej sytuacji w nocy. Dzięki wyjaśnieniom weterynarza okazało się, że pytony właśnie w taki sposób mierzą czy są wystarczająco długie, żeby połknąć swoją ofiarę. : ) Łatwo się domyśleć, dlaczego przestał jeść...*

## Opowieść C

*Smutne. Przypomina mi to historię z bałwankami. Był sobie wredny facet, który rozjeżdżał dzieciom samochodem bałwanki. Co dzieci ulepiły to on rozjeżdżał. Pewnego razu dzieci ulepiły ślicznego bałwanka na hydrancie. Duży im wyszedł ten bałwanek. I Pan Wredotek z impetem spróbował go rozjechać. Ja nie jestem tak subtelny, jak te dzieci. Ja nakarmiłbym człowieka tym, czym on nakarmił psa. I na pewno nie skracałbym mu cierpienia.*

## Opowieść D

*Wczoraj w popołudniowej audycji w trójce prowadzący przeczytał maila od słuchaczki z Poznania: ...na pewnym osiedlu dzieci ulepiły bałwana - ktoś z premedytacją go rozjechał samochodem, następnego dnia ulepiły go jeszcze raz i znów go ktoś rozjechał, trzeciego dnia ulepiły go na hydrancie.... Nie wiem, czy to prawda, ale chciałbym zobaczyć minę kierowcy*



## podobieństwo cosinusowe

wektor dokumentu ( $\vec{V}(d)$ ) wektor, którego składowe odpowiadają wyrazom

## podobieństwo cosinusowe

wektor dokumentu ( $\vec{V}(d)$ ) wektor, którego składowe odpowiadają wyrazom

podobieństwo cosinusowe ( $\text{sim}(d_1, d_2)$ ) cosinus kąta między wektorami dokumentów:

$$\text{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)| |\vec{V}(d_2)|}$$

## podobieństwo cosinusowe

wektor dokumentu ( $\vec{V}(d)$ ) wektor, którego składowe odpowiadają wyrazom

podobieństwo cosinusowe ( $\text{sim}(d_1, d_2)$ ) cosinus kąta między wektorami dokumentów:

$$\text{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)| |\vec{V}(d_2)|}$$

Traktując zapytanie  $q$  jako dokument otrzymamy...

*Cosine score:*

$$\text{Score}(q, d) = \frac{\vec{V}(q) \cdot \vec{V}(d)}{|\vec{V}(q)| |\vec{V}(d)|}$$

## modele języka

**Model języka** przypisuje prawdopodobieństwa napisom.

Jeśli  $M$  ma być modelem języka polskiego, oczekivalibyśmy, że dla napisów:

- $z_1$  *W tym stanie rzeczy pan Ignacy coraz częściej myślał o Wokulskim.*
- $z_2$  *Po wypełniony zbiornik pełny i należne kwotę, usłyszała w attendant*
- $z_3$  *xxxxyzzzzzzit backspace hoooooooooop x y z*

zachodzić będzie:

$$P(z_1|M) > P(z_2|M) > P(z_3|M)$$

## modele języka w wyszukiwaniu informacji

Na podstawie każdego dokumentu  $d$  tworzymy model języka  $M_d$ .

ocena oparta na modelach unigramowych

$$\text{Score}(q, d) = P(q|M_d) = K_q \prod_{t \in q} P(t|M_d)^{\text{tf}_{t,d}}$$